A Comparative Evaluation of Selected Statistical Software for Computing Various Categorical Analyses

Nancy McDermott and Cynthia White
Social Science Computing Cooperative
University of Wisconsin - Madison

Introduction

This paper is a comparative evaluation of statistical software for computing various categorical analyses including logistic regression, multinomial logits, and loglinear analysis. The following statistical packages were included in the evaluation: The SAS System® (version 6.09), STATA (version 3.1), SPSS (version 4.0), GLIM (version 3.77), and LIMDEP (version 6.0).

Large data sets were selected for analysis. The code for the analyses is presented for each of the software packages. Important and unique features of the analyses are noted. Following the output, performance comparisons on a Sparc 10/512 MP running UNIX with 128Mb memory are provided. The UNIX `time` command was used to compare the performances of the statistical packages. The paper also makes some recommendations on the appropriate package to use in certain situations.

Logistic Regression

The data set used for the logistic regression is from the 1980 World Fertility Survey in the Cote d'Ivoire. The data represent responses to interviews of a stratified random sample of women ages 15-50 in the Cote d'Ivoire. A total of 5764 women were interviewed. However, for this analysis, only 4165 married, self-reporting fecund women were included in the analysis.

The dependent variable, WANTYES, was a woman's response when asked whether or not she wanted more children. The independent variables were AGE, number of children ever born (NUMLIV), percent of children who died (PERMORT), EDUCATION, RELIGION, and urbanization (CITY). RELIGION and CITY were categorical variables with three levels each.

Logistic Regression in the SAS System

The SAS System has several procedures which can carry out logistic regression including LOGISTIC, GENMOD, CATMOD, and NLIN. The LOGISTIC procedure was used for this discussion.

The following SAS code requests the logistic regression with WANTYES as the dependent variable. Among the regressors are RELIGD2, RELIGD3, CITYD2, and CITYD3 which are dummy variables created from the RELIGION and CITY categorical variables. Note that RELIGD1 and CITYD1 were not included in the model so that it would not be overdetermined.

```
proc logistic descending;
   model wantyes=age numliv educ permort religd2
                 religd3 cityd2 cityd3
                 / risklimits lackfit ctable;
```

Unlike other statistical packages, by default PROC LOGISTIC models the probability that the event equals zero. To change this to model the probability that the event equals one as in other packages, specify the DESCENDING option on the PROC LOGISTIC statement. If you do not add the DESCENDING option, your parameter estimates may be opposite in sign of what you may get from other statistical packages.

The RISKLIMITS option on the MODEL statement requests confidence intervals for the conditional odds ratio. 95% confidence intervals are computed by default. The LACKFIT option on the MODEL statement requests the Hosmer-Lemeshow Goodness-of-Fit Test. Only one other statistical package (STATA) provided output for these two statistics.

Logistic Regression in STATA

The following STATA code requests the logistic regression:

```
logistic wantyes age numliv educ permort religd2
                 religd3 cityd2 cityd3
lfit, group(10)
logit
```

The LFIT command requests the Hosmer-Lemeshow test for Goodness-of-Fit. Unlike the SAS System, STATA allows you to specify the number of groups to construct for the Hosmer-Lemeshow Goodness-of-Fit test. The SAS System uses approximately 10 groups when constructing the test. STATA reported a Hosmer-Lemeshow Goodness-of-Fit test of 8.42 for this example while the SAS System reported 9.865. The reason for the difference is unknown because both packages constructed the same number of groups after ordering on the predicted probabilities. You get the underlying coefficients for the odds ratios by typing LOGIT without arguments after the LOGISTIC command.

Logistic Regression in SPSS

SPSS has several commands which can carry out logistic regression including LOGISTIC REGRESSION, LOGLINEAR, HILOGLIN, and NLR. The LOGISTIC REGRESSION command was used for this discussion. Note that it is not necessary to create the dummy variables for the two categorical variables (RELIGION and CITY) because SPSS's LOGISTIC REGRESSION procedure generates them automatically. Only one other package (GLIM) offered this feature. The following SPSS code requests the logistic regression:

```
logistic regression wantyes with age numliv educ
                                permort religion city
   /external
   /categorical=religion city
   /contrast(religion)=simple(1)
   /contrast(city)=simple(1)
```

The /EXTERNAL subcommand was used to conserve memory. The /CATEGORICAL subcommand was included to declare the categorical variables. The two /CONTRAST subcommands were used to set the reference category to one so as to make the coefficients comparable to the output from the other statistical packages. There are four other types of contrasts available in SPSS: deviations from the overall effect, difference or reverse Helmert contrasts, Helmert contrasts, and polynomial contrasts.

The LOGISTIC REGRESSION command computes a Goodness-of-Fit test by default. However, this statistic is not appropriate for data like these because there are very few observations at each observed level of the covariate. Hence, the Goodness-of-Fit statistic does not have an approximate chi-squared distribution. A test similar to the Hosmer-Lemeshow Goodness-of-Fit tests computed by the SAS System and STATA would be much more appropriate for the data in this example. SPSS does not compute this test, however.

Logistic Regression in GLIM

GLIM fits generalized linear models, as defined by Nelder and Wedderburn (1972), which include logistic regression. You need to specify a binomial distribution function with a logit link function. The following commands fit the logistic model:

```
$factor relig 3 city 3 $
$yvar wantyes$
$calc n=1$
$error binomial n$
$link g$
$fit age+numliv+ educ+ permort+ relig+ city$
$display e$
```

The FACTOR command defines which variables are categorical. Only one other package (SPSS) can generate the dummy variables automatically. The YVAR command specifies the dependent variable. You must set n equal to 1 with the CALC command because there is only one measurement on each person. The ERROR specification is binomial with the total number of observations on each person equal to 1. The LINK G command specifies that a logit link function will be used for the fit. The FIT command fits the model specified.

Finally, the DISPLAY directive instructs GLIM to display the results of the model fit. In this case, DISPLAY E instructs GLIM to display the parameter estimates and their standard errors, including extrinsically aliased parameters. GLIM does not provide a test similar to the Hosmer-Lemeshow Goodness-of-Fit tests computed by the SAS System and STATA.

Logistic Regression in LIMDEP

The easiest way to carry out a simple logistic regression in LIMDEP is with the LOGIT command. The LOGIT command carries out both binomial and multinomial logit models.

```
logit; lhs=wantyes;
rhs=one,age,numliv,educ,permort,religd2,religd3,
    cityd2,cityd3 $
```

LIMDEP does not provide a test similar to the Hosmer-Lemeshow Goodness-of-Fit tests computed by the SAS System and STATA.

Performance Comparisons for Logistic Regression

Each program for each package was run 10 times. The average time in seconds spent in execution of the program (not real time) is shown in the table below. Times could not be reported for SPSS because the `time` command did not accurately report these for SPSS.

Only a basic logistic regression was specified for each package. In order to make a fair comparison, the Hosmer-Lemeshow Goodness-of-Fit test was not specified because this test can be computationally intensive and thus inflate the times for the two packages that can compute the test.

Package	Time in Seconds
SAS	1.50 (1)
STATA	3.11 (2)
GLIM	6.09 (3)
LIMDEP	31.22 (4)

The number in parentheses represents the package's relative rank for performance.

Recommendations for Logistic Regression

All of the statistical packages considered provided a simple procedure for computing a logistic regression. Unless you need a particular option or CPU time is a factor, it may be more convenient just to use the package with which you are most familiar. SAS's LOGISTIC procedure provided the most options, especially in the areas of criteria for assessing the fit of the model and rank correlation between the observed response and the predicted probabilities. If you only wanted a Goodness-of-Fit test for assessing fit, either the SAS System or STATA would be a good choice. Both packages compute the Hosmer-Lemeshow test. Also, the SAS System and STATA were the only packages that computed odds ratios with confidence intervals. SPSS computed the odds ratio but without confidence intervals.

SPSS provided most of the options that the SAS System and STATA did. One nice feature offered by SPSS that only one other package (GLIM) offered was the automatic construction of dummy variables for categorical independent variables. In addition, it provides five different types of contrasts for the categorical variables used for interpreting the coefficients in different ways.

The SAS System and STATA ranked highest in execution time. STATA's good performance was not unexpected because STATA puts all the data in memory instead of using swap space. Although this method of execution can put a huge drain on a machine's memory when a large job is executing, it usually means the package will execute jobs very quickly. What was unexpected was that the SAS System actually performed better than STATA for the larger analysis. The SAS System does make use of swap space instead of putting everything in memory. Both the SAS System and STATA appear to be good choices when CPU time is a factor. Because of slow performance, you may want to avoid LIMDEP for large problems.

Multinomial Logit Regression

SAS, STATA, and LIMDEP were the only packages compared for the multinomial logit runs. SPSS and GLIM were not included because they do not offer a multinomial procedure. Although not considered in this paper, for multinomial models that have an equivalent loglinear model, GLIM or SPSS's LOGLINEAR procedure could be used to fit these models.

The data set used for the multinomial regression was an extract based on the 5%Public Use Microdata Survey (PUMS). The variables include five occupation/industry categories, age in years, educational attainment in years, sex with two categories, race with two categories, and time with two categories (1980 or 1990). The variable representing the five occupation/industry categories was used as the dependent variable. There were 28,369 observations in the extract. A full five-way model was fit which required that 128 parameters be fit.

Multinomial Logit Regression in the SAS System

The CATMOD procedure can be used to fit multinomial models in the SAS System. This procedure fits linear models to functions of response frequencies and uses either maximum-likelihood estimation or weighted least squares estimation. The following SAS statements fit the 5-way model:

```
proc catmod;
      direct age yearsch;
      model occup=sex|race|age|yearsch|year
      / ml nogls noprofile;
```

PROC CATMOD generates the design matrix for categorical explanatory variables automatically. The SAS System was the only software package examined that had this feature. In PROC CATMOD, explanatory variables are assumed to be categorical unless declared otherwise with a DIRECT statement. The ML and NOGLS options instruct the SAS System to compute maximum-likelihood estimates instead of weighted-least-squares estimates.

PROC CATMOD uses a different parameterization for the explanatory variables than was used for STATA and LIMDEP. PROC CATMOD constrains the parameters to sum to zero. In other words, PROC CATMOD uses a full-rank center-point parameterization to build design matrices. For example, when the race variable, BLACK, is specified as a categorical variable, each value gets coded internally as either 1 or -1 instead of 1 or 0 as was done by the other packages. The sum-to-zero constraint requires that the last level of an effect be the negative of the sum of the other levels of the effect.

If you do not want the full-rank center-point parameterization that PROC CATMOD uses, you can construct the indicator variables yourself in the data step and then insert a DIRECT statement which instructs PROC CATMOD to treat the variables specified as quantitative rather than qualitative. No matter which way you choose to specify the model, you will get a solution to the same underlying model along with the same predicted probabilities.

Multinomial Logit Regression in STATA

The MLOGIT command can be used to fit a multinomial model in STATA. The maximum number of explanatory variables that can be fit in any of STATA's estimation procedures is 400.

Following is the MLOGIT command to fit the one-way model:

```
mlogit occup x1-x31
mlogit, rrr
```

The MLOGIT command does not generate the indicator variables corresponding to the explanatory variables automatically. These were generated using the GLMMOD procedure in the SAS System. The MLOGIT,RRR command was used to display the estimated coefficients transformed to relative risk ratios (e^b rather than b). STATA was the only package to provide this output.

Multinomial Logit Regression in LIMDEP

The LOGIT command in LIMDEP fits both logit models and multinomial models. The maximum number of parameters that can be estimated in a model in LOGIT is 150. Keep in mind that the total number of parameters is the product of the number of explanatory variables and the number of levels of the outcome variable minus one.

Following is the LOGIT command to fit the one-way model:

```
logit; lhs=occup;
rhs=one,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,
x13,x14,x15,x16,x17,x18,x19,x20,x21,x22,x23,x24,
x25,x26,x27,x28,x29,x30,x31 $
```

Like with the MLOGIT command in STATA, LOGIT will not generate indicator variables for you. You must create them manually.

Performance Comparisons for Multinomial Logit Regression

Each program for each package was run 10 times. The average time in minutes spent in execution of the program is shown in the table below.

Package	Time in Minutes
SAS	1.97 (1)
STATA	28.68 (2)
LIMDEP	218.54 (3)

The number in parentheses represents the package's relative rank for performance.

Recommendations for Multinomial Logit Regression

The complexity and number of effects in the model will probably be the deciding factors in your decision about which package to use. For complex models, the multinomial procedures provided by LIMDEP and STATA are not good choices because they do not construct the indicator variables for the design matrix automatically. For models with many effects, you will also want to avoid LIMDEP and STATA because of their restriction on the number of parameters that can be fit. LIMDEP has a limit of 150 parameters and STATA has a limit of 400.

The SAS System was also the clear winner when it came to performance. You could still compute a large model in a

reasonable amount of time with STATA but you would definitely want to avoid LIMDEP for large models.

Loglinear Analysis

The data set used for this analysis is a 36x2x2x4x2 contingency table based on the 5% Public Use Microdata Survey (PUMS). The variables include 36 occupation/industry categories, nativity with two categories, sex with two categories, ethnicity with four categories, and time with two categories (1980 or 1990). The table has a total of 1152 cells.

For each package, an attempt was made to fit a sequence of models, beginning with the saturated model and continuing through models of decreased complexity, until a model can be fit without running out of memory. The largest model the package can compute is reported. Then, a comparison of times between each of the packages is made for two models, one simple and one more complex.

Loglinear Models in the SAS System

The SAS System has two procedures which carry out a loglinear analysis: CATMOD and GENMOD. The CATMOD procedure fits linear models to functions of response frequencies. The GENMOD procedure was recently introduced in version 6.09 of the SAS System and fits generalized linear models as defined by Nelder and Wedderburn (1972). PROC GENMOD was used for this discussion.

You analyze this 36x2x2x4x2 contingency table by means of a generalized linear model with a log link function. The error distribution appropriate to the counts is Poisson. You specify the link function and error distribution in SAS's GENMOD procedure with the DIST=POISSON and LINK=LOG options on the MODEL statement. The following SAS code fits the one-way model:

```
proc genmod;
   class ethnic sex native occup yr;
   model count=yr occup ethnic sex native
      / dist=poisson link=log;
```

ETHNIC, SEX, NATIVE, OCCUP, and YR are specified as CLASS variables so that PROC GENMOD automatically generates the indicator variables associated with these variables.

Loglinear Models in STATA

STATA has two commands for carrying out a loglinear analysis: LOGLIN and POISSON. LOGLIN is a user-contributed program which is available as a STATA "ado" file. Both LOGLIN and POISSON estimate a Poisson maximum-likelihood regression for the number of occurrences of an event. The main difference between the two commands is in the generation of indicator variables. Indicator variables are not generated automatically with the POISSON command as they are in the LOGLIN command. However, with the LOGLIN command, you are restricted to four effects. Since the data set for this example contains five effects, only the POISSON command was used. The indicator variables were generated using the GLMMOD procedure in the SAS System.

The following STATA code fits the one-way model:

```
poisson count yr occupd2-occupd36 ethnicd2-
 ethnicd4 sex native
```

Loglinear Models in SPSS

You can analyze loglinear models in SPSS using either the HILOGLINEAR or the LOGLINEAR procedure. Both of these procedures have slightly different features and the features you need will determine which procedure you use. HILOGLINEAR is well suited for hierarchical log-linear models in which models are nested one within the other. It may be best to use HILOGLINEAR when the intent is to select the best possible model. The design statement syntax of HILOGLINEAR is somewhat less complicated than the design syntax of LOGLINEAR. In HILOGLINEAR, lower order interaction terms will automatically be included in the design if the highest order interaction term is specified in the design statement. One drawback of HILOGLINEAR is that it will only produce parameter estimates for the saturated model. LOGLINEAR, on the other hand, will produce estimates for all models. LOGLINEAR was used for this paper.

The following set of commands were used to read in the data and compute the one-way model using LOGLINEAR:

```
weight by count
loglinear ethnic(1,4) yr(1,2) sex(1,2)
native(1,2) occup(1,36)
   /print=estim
   /design=ethnic yr sex native occup
```

The statement WEIGHT BY COUNT instructs LOGLINEAR to use the counts from the table as weights in the estimation of the loglinear model. On the LOGLINEAR command statement, you must specify the levels of each categorical variable that will be used in the model. SPSS automatically generates the indicator variables associated with these variables. Lastly, you must specify a DESIGN subcommand. In this case, the /DESIGN subcommand specifies that LOGLINEAR should fit a one-way model.

The parameter estimates are different from those of the other three packages. This is because SPSS parameterizes the model differently. By default, SPSS computes the parameter estimates by constructing contrasts of the deviation from the overall effect. All the other packages examined computed the parameter estimates by constructing contrasts for each level of a factor to the last level. The CONTRAST subcommand in SPSS's LOGLINEAR command can be used to specify other types of contrasts including that used by other packages, but only for models that do not contain interaction effects. No matter which way you choose to specify the model, you will get a solution to the same underlying model along with the same predicted probabilities.

Loglinear Models in GLIM

GLIM fits generalized linear models, as defined by Nelder and Wedderburn (1972). To analyze contingency tables by means of a generalized linear model with GLIM, you specify a log link function and a Poisson error distribution.

The following set of commands were used to read in the data and compute the one-way model using GLIM:

```
$units 1152
$factor ethnic 4 yr 2 sex 2 native 2 occup 36
$data yr ethnic sex native occup count
$dinput 7
$yvar count
$error pois
$fit yr ethnic sex native occup
$display e
$look %X2
```

The FACTORS directive identifies the explanatory variables and instructs GLIM to generate the indicator variables associated with these variables. The YVAR directive specifies the dependent variable containing the cell counts. The ERROR specification is Poisson. No LINK directive was specified because the log link is used by default with the Poisson error. The FIT directive fits the one-way loglinear model. The DISPLAY directive instructs GLIM to display the parameter estimates. The LOOK directive is used to display the Pearson's chi-square statistic which is helpful is assessing the goodness-of-fit of a given model.

Loglinear Models in LIMDEP

LIMDEP uses a Poisson regression model to fit loglinear models. As is true with the POISSON command in STATA, the POISSON command in LIMDEP will not generate indicator variables for you. You must create them manually.

Following is the POISSON command used to perform the loglinear analysis and the results:

```
poisson; lhs=count;
rhs=one,yr,occupd2,...occupd36,
ethnicd2,ethnicd3,ethnicd4,sex, native$
```

Performance Comparisons for the Loglinear Model

For each package, an attempt was made to fit a sequence of models, beginning with a saturated model and continuing through models of decreased complexity, until a model could be fit without running out of memory.

STATA and LIMDEP could not be included in several of these performance comparisons because they have limits on the number of variables or effects in a given model. LIMDEP has a limit of 200 variables in a data set and STATA limits the number of effects in a model to 400. The saturated model, the model including all four-way interactions, and the model including all three-way interactions all have over 400 effects and thus could not be included in the comparisons for STATA. It could only be included in the comparison of the model with all two-way interactions. Even the two-way interaction model has over 200 variables in the design matrix so LIMDEP was not included in any of the performance comparisons.

Only GLIM could fit the saturated model or the model involving all four-way interactions. SPSS and the SAS System required more memory than was available. Next, a fit was attempted for the model involving three-way interactions. The model was run for the SAS System and SPSS. Both packages were able to fit this model without running out of memory.

Two models were used for the CPU comparisons: the model with all two-way interactions and the model with all three-way interactions. The program for each package was run 10 times for the smaller model and three times for the larger model. The average time in minutes spent in execution of the program (not real time) is shown in the table below. Times could not be reported for SPSS because the `time` command did not accurately report these for SPSS.

	Time in Minutes Spent in Execution of the Program	
	2-Way Model:	**3-Way Model:**
SAS	0.34 (1)	12.29 (1)
STATA	2.50 (3)	More effects than allowed
GLIM	1.40 (2)	15.79 (2)

The number in parentheses represents the package's relative rank for performance.

Even though the time spent in execution of the program was not reported accurately for SPSS, it is still possible to get a rough idea of how SPSS compared to the other packages, if you look at the average results of the real time that elapsed during execution of the program. The real time for the SPSS programs was much longer than for the other programs. For example, the average (over three runs) real time for the larger model for SPSS was 255.24 minutes, 27.70 minutes for GLIM, and 20.96 minutes for the SAS System.

Recommendations for the Loglinear Model

The complexity and number of effects in the model will probably be the deciding factors in your decision about which package to use. For complex models, the POISSON commands in LIMDEP and STATA are not a good choice because they do not construct the indicator variables for the design matrix automatically. And, even though STATA's LOGLIN command will construct the design matrix for you, it has the four factor restriction.

You will also want to avoid LIMDEP and STATA for models with many effects because of their restriction on the number of variables that can be used. LIMDEP has a limit of 200 variables in a data set and STATA limits the number of effects in a model to 400.

In terms of memory requirements, by far, GLIM required the least amount of memory to compute a loglinear model. GLIM could fit the saturated model for the 36×2×2×4×2 contingency table whereas the largest model that the SAS System and SPSS could fit was the model involving all three-way interactions.

If you have a fairly large model, you might also want to avoid SPSS if CPU usage is a concern. It was much slower than the SAS System and GLIM. The SAS System and GLIM performed about equally well in this category.

GLIM was the clear winner when it came to performance, but, it also offers the fewest options for enhancing your output. The SAS System and SPSS's loglinear procedures provided lots of useful output that is not available in GLIM or STATA.

In summary, for large models, GLIM may be the only package that can fit the model without running out of memory. STATA and LIMDEP, with their restrictions on the number of effects, are clear losers. If you do not want the hassle of generating your own design matrix, avoid LIMDEP and the POISSON command in STATA. For moderate sized models, the SAS System provides an easy-to-use procedure with lots of useful options. For smaller models, it may be more convenient to use the package with which you are most familiar unless you need a particular option.

References

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley and Sons, Inc.

Baker, R.J. and J.A. Nelder (1987), *The GLIM System Release 3.77 Manual - Edition 2*, Numerical Algorithms Group Inc., Downers Grove, IL.

Fienberg, S. E. (1977), *The Analysis of Cross-Classified Categorical Data*, MIT Press.

Greene, W. H. (1992), *Limdep User's Manual and Reference Guide: Version 6*, Bellport, NY: Econometric Software, Inc.

Hosmer, D.W., Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley and Sons, Inc.

McDermott, N.J. (1995), "A Comparative Evaluation of Selected Statistical Software for Computing Multinomial Logit Models," Center for Demography Working Paper Series, #95-01, University of Wisconsin-Madison.

McDermott, N.J. and C. White (1994), "A Comparative Evaluation of Selected Statistical Software for Computing a Logistic Regression," Center for Demography Working Paper Series, #94-27, University of Wisconsin-Madison.

McDermott, N.J. and C. White (1994), "A Comparative Evaluation of Selected Statistical Software for Computing Loglinear Models," Center for Demography Working Paper Series, #94-28, University of Wisconsin-Madison.

Nelder, J. A. and R.W.M. Wedderburn (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society A*, 135, 370:384.

Norusis, M. J. (1990), *Advanced Statistics User's Guide*, Chicago, IL: SPSS Inc.

SAS Institute Inc. (1992), *SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1989), *SAS/STAT Users Guide, Version 6, Fourth Edition, Volume 1*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1989), *SAS/STAT Users Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc.

SPSS Inc. (1990), *SPSS Reference Guide*, Chicago, IL: SPSS Inc.

Stata Corporation (1993), *Stata Reference Manual: Release 3.1*, College Station, TX.

Author

A much more detailed account of this work including output from each of the computer runs may be found in the three papers listed in the References section of this paper by McDermott. These papers may be requested from the address listed below. You are welcome to address questions or comments there as well.

Nancy J. McDermott
Social Science Computing Cooperative
1180 Observatory Drive
University of Wisconsin
Madison, WI 53706

Internet Address: mcdermot@ssc.wisc.edu